Supplementary Methods

1. Steps of PEP_scaffolder

The main steps of the PEP_scaffolder algorithm are outlined in the Supplementary Fig. S1.

(1) Aligning and selecting guides

The alignments of proteins to contigs by BLAT are subjected into scaffolding. Protein sequences are aligned to contigs using BLAT with the parameters of '-t=dnax -q=prot'. The percent identity (PI) of one alignment between one protein and one contig is calculated using the web-based BLAT percent identity formula in the UCSC Genome Browser (http://genome.ucsc.edu/FAQ/FAQblat.html#blat4). If the percent identities of alignments are over a certain minimal percent identity (MPI), these alignments are retained to ensure the alignment reliability. The protein length coverage (LC) in one retained alignment is calculated as follows:

Length coverage = aligned length / total protein length

The alignments under a certain minimal length coverage (MLC) indicate that the proteins are not fully covered. The proteins not fully covered in all alignments are used as 'guides', with the whole protein sequences divided into different fragments. The guides and corresponding alignments are kept for the downstream scaffolding.

(2) Clustering alignments into blocks and ordering the blocks

For each guide, all alignments are ordered based on their query start positions. If alignments have common query regions in guide, they are clustered into one block. Then the longest query region is selected to represent this block.

All blocks are ordered based on their coordinates in the guide to re-build the guide. Since each block corresponds to only one contig, all contigs are sorted following the order of the blocks.

(3) Filtering erroneous connections with large introns

After the step of (2), two contigs constitute one connection, where the first contig is considered as the donator and the second sequence the acceptor. This guide is considered as supporting evidence of this connection.

In this connection, the DNA sequences between two adjacent blocks might be an intron. Assuming that two neighboring blocks (***a*** and ***b***) are located in two contigs (***A*** and ***B***), respectively, the variable of ***L(a, b)*** is defined as the possible intron length (IL) between two blocks. The variable is calculated as follows:

$$L(a, b) \geq [Length(A) - end(a) + start(b)]$$

where **Length (A)** is the length of contig ***A***, **end (a)** is the end position of **block *a*** in contig ***A*** and **start (b)** is the start position of **block *b*** in contig ***B***. If ***L(a,b)*** is over a certain maximal intron length (MIL), then this connection is filtered out because an extremely large intron is likely a result of misalignment.

(4) Finding the optimal connection for each sequence

Among all retained connections, a sequence could be a donator and/or an acceptor in different connections. Considering that one contig might have many donators and/or acceptors, we find an optimal donator and or acceptor for it.

For each donator (acceptor) with many acceptors (donators), the connection with the maximal number of supporting guides is retained as the optimal connection for it. If one donator (acceptor) has two or more acceptors (donators) with the same number of supporting evidence, this donator (acceptor) is considered to have no acceptor (donator) and all connections of this donator (acceptor) are discarded. Highly similar homologs might result in two exonic genomic fragments from the homologs being connected together. The process of finding the optimal connection for each contig decreases the influence of homologs during the scaffolding.

(5) Building paths by walking the optimal connections

The reserved contig is assigned to at most two connections and classified into three types: (i) crossover point where the sequence is the donator in one connection and the acceptor in the other connection; (ii) donator in only one connection; and (iii) acceptor in only one connection.

For each donator, we search for its optimal crossover point and search for a new crossover point for the prior crossover point. Repeat these searches to extend the scaffolding path to one acceptor. After all the donators are walked, all contigs are attributed into scaffolding paths.

(6) Estimating gap size from intron size distribution

The gap between two connected contigs is from an intron. To estimate the gap size of two adjacent contigs, we plot the intron size distribution from the proteins fully covered in the genome and estimate the median intron size. Then, if ***L(a, b)*** is smaller than the median size, we insert a sequence composed of letter ‘N’, the number of which is the difference between the median intron size and ***L(a, b)***. Otherwise, 100 Ns are inserted between two contigs to indicate a possible gap.

2. Estimating scaffolding accuracy

Following the Genome Assembly Gold Standard Evaluations pipeline (Salzberg, *et al.*, 2012), we classified the connections into five categories using hg38 assembly as reference and measured the accuracy of PEP_scaffolder. Assuming that hg38 assembly is totally correct, all connections could be tallied into five types. (i) Consistence where two connected sequences by PEP_scaffolder have the same order and orientation as hg38 assembly. (ii) Inversions where two scaffolded sequences have the same order as hg38 assembly but the orientation of one sequence is different from the reference. (iii) Correctable relocations where two distant contigs from the same chromosome are merged together with an interval between them smaller than MIL. (iv) If two distant contigs in one chromosome have an interval larger than MIL and are scaffolded together, this connection is considered as an erroneous relocation. (v) Translocation where two sequences from two chromosomes are linked. The links of inversion, erroneous relocation and translocation are considered to be scaffolding errors.

After all connections are classified, the scaffolding accuracy is calculated as the ratio of (consistence + correctable relocations) / total connections.

3. Calculating genome coverage and N50 size

The scaffolding performance is affected by genome coverage of aligned proteins. Genome coverage is measured as the ratio of the total length of all the bases covered in the protein alignment regions to the total base number of the genome (Sims, et al., 2014). The genomic bases covered in the alignments include the intronic bases and exonic bases. Note that the alignment regions only include the full-covered proteins and guiders.

The N50 size and N50 number are used as metrics to determine the scaffolding performance, without consideration of scaffolding errors. The N50 length is the length x such that 50% of the genome size is contained in sequences of length x or greater. The N50 number is the number of sequences with lengths greater than the N50 length. This strategy is commonly adopted for a new genome assembly. Furthermore, to produce a revised and accurate picture of the assembly, we follow the Genome Assembly Gold Standard Evaluations pipeline

(Salzberg, et al., 2012) and measure the corrected N50 size considering scaffolding errors (including inversions, erroneous relocations and translocations). We split scaffolds at every error point and compute the corrected N50 length.

More genome regions covered by proteins would generate longer scaffolds. Using 15 scaffolding results in Supplementary Tables S1, S2 and S3, we calculate the correlation coefficient (*R*) between N50 size improvement and genome coverage. Using *t*-test, we investigate whether the *R* is statistically different from zero, which is what one would expect by chance.

4. Estimating the proportion of fully covered genes

Using BLAT (Kent, 2002), we aligned human Swiss-Prot proteins to three assemblies (the contigs, the PEP_scaffolder assembly and the hg38 assembly). If the alignment of one protein had a coverage over 90%, we considered this protein to be fully covered and complete. Then we calculated the proportion of fully covered proteins among all proteins.

5. Recipes for fly genome scaffolding

Here we describe how we run each algorithm with fly Ensembl proteins to scaffold fly contigs in this study. Fly genome and proteins were downloaded from Ensembl database. Then the genome was fragmented into contigs of same length of 10 kb. Each scaffolder was run with the default parameters. We placed the genome sequence, proteins, and all the scaffolded connections at our website (http://www.fishbrowser.org/software/PEP_scaffolder). The files of fly contigs and proteins are named as ‘fly_contig.fasta’ and ‘fly_Ensembl_protein.fasta’. We ran each scaffolder with one processor on the same machine. These recipes allow others to replicate the comparison.

Because ESPRIT only produced the connections without genome sequences generated, we compared the correctness of connections of three scaffolders. After scaffolding, we followed the above strategy to evaluate the accuracy of each method.

(1) PEP_scaffolder

```
## align proteins to target genome
blat –t=dnax –q=prot fly_contig.fasta fly_Ensembl_protein.fasta fly.psl -noHead

## running PEP_scaffolder
sh PEP_scaffolder –d ./ –i fly.psl –j fly_contig.fasta
## After scaffolding, the file of “final.nodes” stores the contig connections.
```

(2) ESPRIT

```
## predict protein-coding genes of target genome.
augustus --species=fly fly_contig.fasta >output.gff

## retrieve predicted protein-coding genes from gff file
perl augustus2fasta.pl output.gff >augustus.fa
(the script ‘augustus2fasta.pl’ is an in-house perl script to retrieve fasta sequences)

##set up a working sub-directory under OMA directory
mkdir myWorkingDir
cd myWorkingDir
```

```
mkdir DB
cp fly_Ensembl_protein.fasta DB/fly2.fa
cp augustus.fa DB/fly.fa
```

##change the parameters

```
cp ../parameters.drw ./
vi parameters.drw
UseEsprit := true
```

## run ESPRIT

```
OMA
```

(After OMA is finished, a sub-directory of “EspritOutput” is generated, where the file of “hits.txt” stores the connections between two predicted proteins)

##generate the genomic connections

```
awk '{if($2=="AUGUSTUS" && $3=="gene") print $1"\t"$4"\t"$5"\t"$7"\t"$9}' output.gff >gene.location
perl add-location.pl gene.location ./EspritOutput/hits.txt >connection.txt
```

(the script ‘add-location.pl’ is an in-house perl script to add the gene location information to connections of predicted proteins and generate the genomic connections)

(3) SWiPS

##format the protein sequences

```
mkdir data
mkdir run
cd scripts
python reformat_fasta.py fly_Ensembl_protein.fasta drosophila_seed.reformated.fa
mv drosophila_seed.reformated.fa ../data
```

## blast format the genome sequences

```
mv fly_contig.fasta ../data/dmel.scafSeq
formatdb -i ../data/dmel.scafSeq -p F
```

## align proteins against contigs

```
python map_tblastn.py    ../data/drosophila_seed.reformated.fa ../data/dmel.scafSeq ../run/tblastn.out 0 100 1e-05 log_tblastn
```

## parse the tblastn alignments

```
python parse_blast.py ../run/tblastn.out ../run/tblastn_all.parsed log
```

##run GeneWise

```
mkdir ../run/genewise/
python prepare_hfinder.py ../run/tblastn_all.parsed ../data/drosophila_seed.reformated.fa../data/dmel.scafSeq 0
100 ../run/gw_0.sh ../run/genewise/
sh ../run/gw_0.sh > ../run/gw_1.out
cat gw_1.out | xargs cat >gw.out
python remap_genewise.py ../data/drosophila_seed.reformated.fa ../run/gw.out ../run/remapped_gw.txt
```

##generate connections

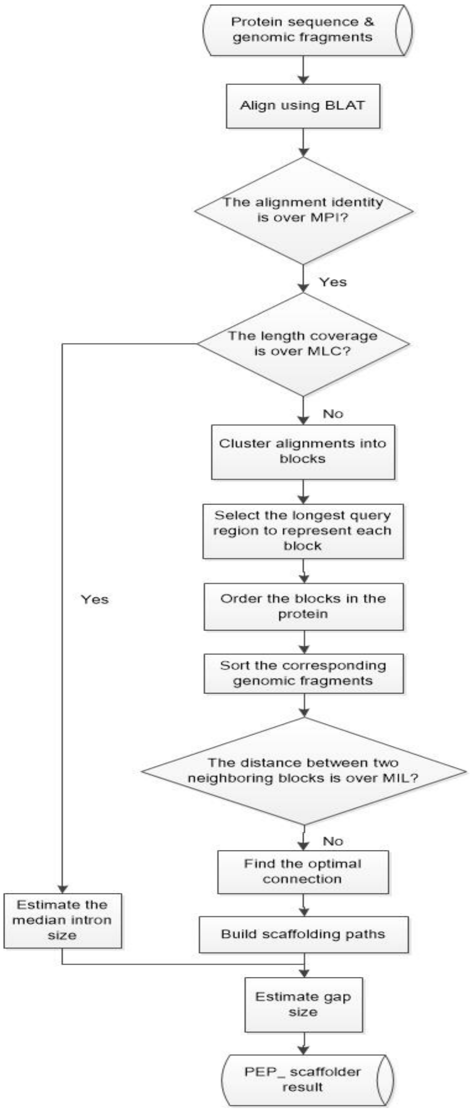
```
python scaffold_contigs.py ../run/remapped_gw.txt ../run/test
```

(After this step is finished, a file of "test.scaff_dir" is generated which stores the scaffolding paths)

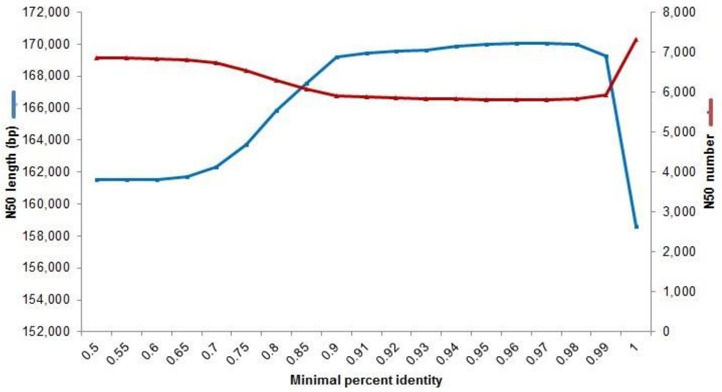
6. Improving the scaffolding accuracy with homologous proteins

Protein sequence variations between species might lead to erroneous scaffolding. To improve the scaffolding accuracy on target species with homologous proteins, we increase the supporting protein number from one to three and measure the scaffolding performance and accuracy.

Supplementary Figure S1. Flow chart of the PEP_scaffolder

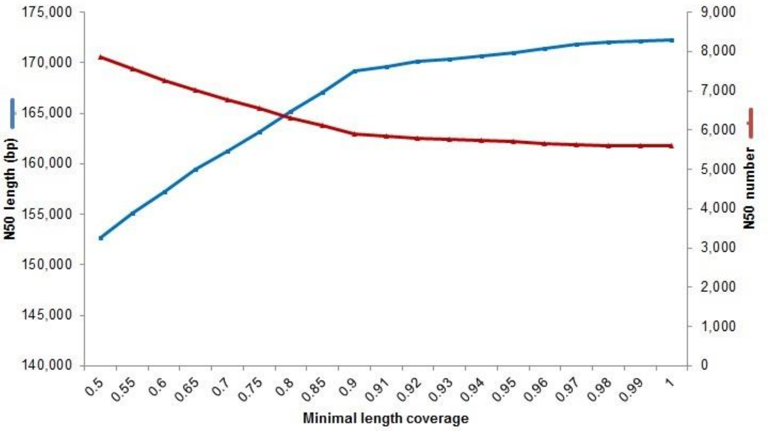

Protein sequence & genomic fragments
Align using BLAT
The alignment identity is over MPI?
Yes
The length coverage is over MLC?
Yes
No
Cluster alignments into blocks
Select the longest query region to represent each block
Order the blocks in the protein
Sort the corresponding genomic fragments
The distance between two neighboring blocks is over MIL?
No
Find the optimal connection
Estimate the median intron size
Build scaffolding paths
Estimate gap size
PEP_ scaffolder result

Supplementary Figure S2. Influence of MPI on the performance of PEP_scaffolder



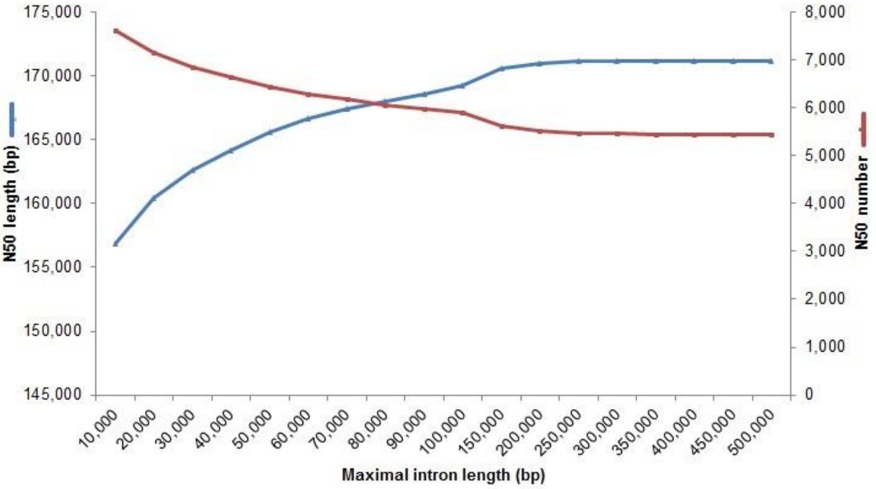
With MLC set as 0.9 and MIL as 150 kb, the N50 length (blue line) is almost up to the saturation when the MPI is over 0.9 but dramatically decreases when the MPI is over 0.99. The N50 number (red line) exhibits the opposite trend to the N50 length.

Supplementary Figure S3. Influence of MLC on the performance of PEP_scaffolder



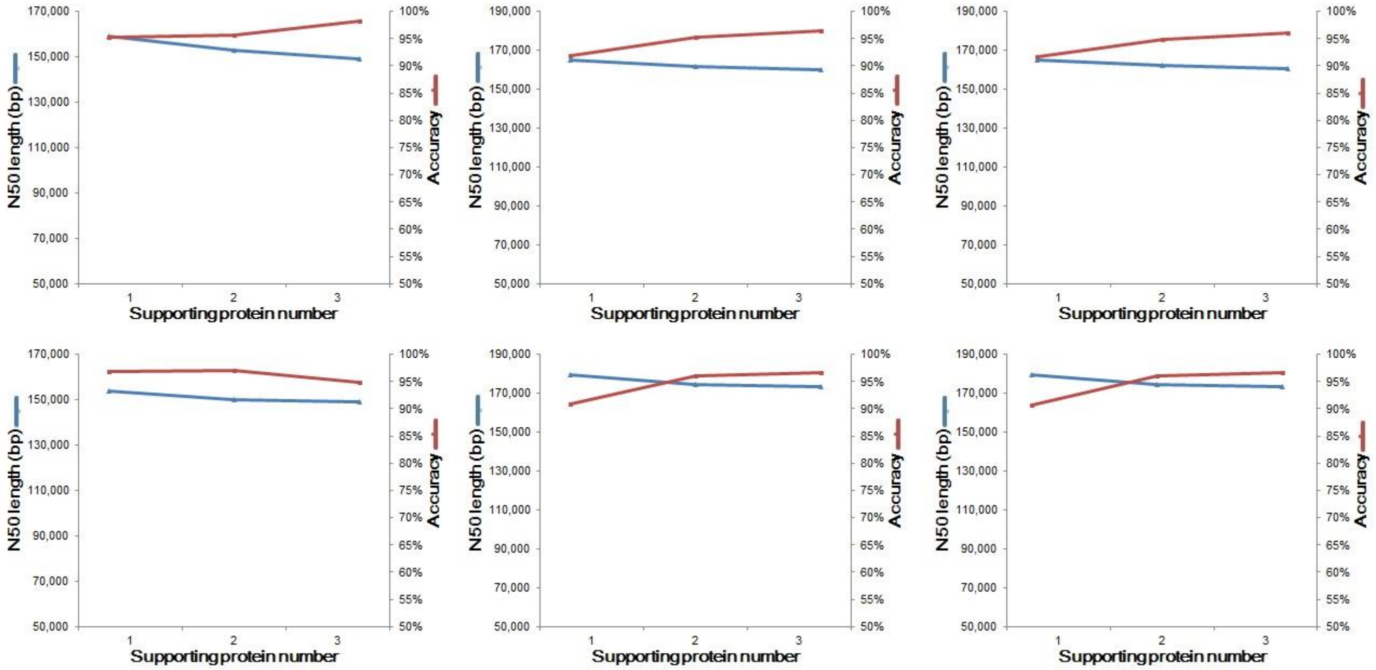
With MIL of 150 kb and MPI of 0.9, the N50 length (blue line) increases to the saturation when the MLC is over 0.9. The N50 number (red line) exhibits the opposite trend to the N50 length.

Supplementary Figure S4. Influence of MIL on the performance of PEP_scaffolder



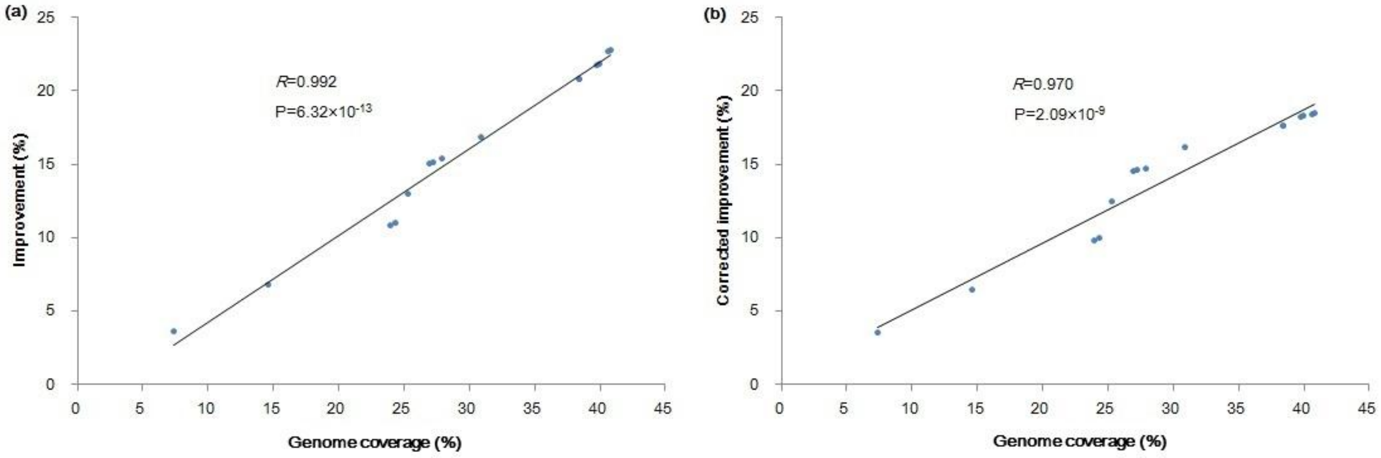
With MPI set as 0.9 and MLC as 0.9, the N50 length (blue line) reaches the saturation point when the MLC is over 150 kb. The red line represents the N50 number.

Supplementary Figure S5. Influence of supporting protein number on the performance of PEP_scaffolder with homologous proteins



The parameters of PEP_scaffolder are set to be optimal. The red line and blue line are the accuracy and the N50 length, respectively. In general, the accuracy of PEP_scaffolder with homologous proteins increases along with the supporting protein number. (a). Rodent Swiss-Prot proteins as ‘guides’. (b). Rodent TrEMBL proteins as ‘guides’. (c). Rodent Swiss-Prot and TrEMBL proteins as ‘guides’. (d). Mammal Swiss-Prot proteins as ‘guides’. (e). Mammal TrEMBL proteins as ‘guides’. (f). Mammal Swiss-Prot and TrEMBL proteins as ‘guides’.

Supplementary Figure S6. The correlation between N50 size improvement and genome coverage



R is the correlation coefficient between (corrected) N50 size improvement and genome coverage. P is the statistical value of *t*-test. (a). The correlation between N50 size improvements and genome coverages using 15 scaffolding results. (b). The correlation between corrected N50 size improvements and genome coverages using 15 scaffolding results.

Supplementary Table S1. The scaffolding performance of PEP_scaffolder with Swiss-Prot proteins on human contigs

	Human	Rodents	Mammals	Hm*	Hmr#
protein number	20,207	26,337	19,830	40,037	66,374
Accuracy					
Consistence	4,912	2,241	1,250	4,937	4,972
Inversions	0	0	0	0	0
Correctable relocations	336	195	89	337	343
Erroneous relocations	134	84	28	136	142
Translocations	9	35	15	19	34
Total correct links$	5,248	2,436	1,339	5,274	5,315
Total links	5,391	2,555	1,382	5,429	5,491
Accuracy	97.35%	95.34%	96.89%	97.14%	96.79%
Length					
N50 size (bp)	171,032	158,880	154,078	171,259	171,583
Improvement	15.01%	6.84%	3.61%	15.16%	15.38%
Inserted gap length (bp)	607,676	282,184	153,611	614,531	621,287
Corrected N50 size (bp)	170,366	158,381	153,942	170,491	170,639
Corrected improvement	14.56%	6.50%	3.51%	14.64%	14.74%
Sequence Number					
PEP_scaffolder sequence number	31,046	33,882	35,054	31,007	30,945
Initial contigs number	36,437	36,437	36,437	36,437	36,437
Coverage					
Covered genome regions (bp)	864,026,701	468,951,125	234,771,636	873,776,724	894,360,984
Genome coverage	26.92%	14.61%	7.32%	27.23%	27.87%

*Hm: Swiss-Prot proteins of human and mammal

#Hmr: Swiss-Prot proteins of human, mammal and rodent

$ The correct links are links of consistence and correctable relocations.

Supplementary Table S2. The scaffolding performance of PEP_scaffolder with TrEMBL proteins on human contigs

	Human	Rodent	Mammal	Hm*	Hmr#
protein number	126,454	245,600	929,543	1,055,997	1,301,597
Accuracy					
Consistence	4,263	3,370	5,680	5,832	5,870
Inversions	0	1	1	1	1
Correctable relocations	297	282	427	436	445
Erroneous relocations	130	136	208	220	234
Translocations	36	189	408	417	504
Total correct links	4,560	3,652	6,107	6,268	6,315
Total links	4,726	3,978	6,724	6,906	7,054
Accuracy	96.49%	91.80%	90.82%	90.76%	89.52%
Length					
N50 size (bp)	168,047	164,842	179,712	180,999	182,433
Improvement	13.00%	10.84%	20.84%	21.71%	22.67%
Inserted gap length (bp)	504,764	437,533	770,606	765,479	782,312
Corrected N50 size (bp)	167,202	163,331	174,933	175,862	176,097
Corrected improvement	12.43%	9.83%	17.63%	18.25%	18.41%
Sequence Number					
PEP_scaffolder sequence number	31,711	32,459	29,705	29,523	29,375

Initial contigs number	36,437	36,437	36,437	36,437	36,437
Coverage					
Covered genome regions (bp)	810,324,216	767,275,122	1,232,641,905	1,275,745,202	1,303,352,495
Genome coverage	25.25%	23.91%	38.41%	39.75%	40.61%

[*] Hm: TrEMBL proteins of human and mammal

[#] Hmr: TrEMBL proteins of human, mammal and rodent

Supplementary Table S3. The scaffolding performance of PEP_scaffolder with Swiss-Prot proteins and TrEMBL proteins on human contigs

	Human[*]	Rodent[$]	Mammal[&]	Hm[%]	Hmr[#]
protein number	146,661	271,937	949,373	1,096,034	1,367,971
Accuracy					
Consistence	5,320	3,396	5,680	5,844	5,885
Inversions	0	1	1	1	1
Correctable relocations	371	284	428	437	446
Erroneous relocations	163	140	208	220	233
Translocations	33	195	411	420	504
Total correct links	5,691	3,680	6,108	6,281	6,331
Total links	5,887	4,016	6,728	6,922	7,069
Accuracy	96.67%	91.63%	90.78%	90.74%	89.56%
Length					
N50 size (bp)	173,729	165,028	179,714	181,140	182,560
Improvement	16.82%	10.97%	20.84%	21.80%	22.76%
Inserted gap length (bp)	636,346	440,726	770,396	768,577	784,642
Corrected N50 size (bp)	172,705	163,504	174,940	175,943	176,257
Corrected improvement	16.13%	9.94%	17.63%	18.31%	18.52%
Sequence Number					
PEP_scaffolder sequence number	30,550	32,421	29,700	29,506	29,359
Initial contigs number	36,437	36,437	36,437	36,437	36,437
Coverage					
Covered genome regions (bp)	990,542,553	780,366,022	1,233,145,894	1,280,744,807	1,309,082,118
Genome coverage	30.86%	24.32%	38.42%	39.91%	40.79%

[*] Human: Swiss-Prot and TrEMBL proteins of human

[$] Rodent: Swiss-Prot and TrEMBL proteins of rodent

[&] Mammal: Swiss-Prot and TrEMBL proteins of mammal

[%] Hm: Swiss-Prot and TrEMBL proteins of human and mammal

[#] Hmr: Swiss-Prot and TrEMBL proteins of human, mammal and rodent

Supplementary Table S4. The scaffolding performance of PEP_scaffolder with Ensembl proteins on fly contigs

	PEP_scaffolder	ESPRIT	SWiPS
Accuracy			
Consistence	3,749	413	1,448
Inversions	0	0	0
Correctable relocations	423	23	243
Erroneous relocations	16	0	13
Translocations	3	5	33
Total correct links	4,172	436	1,737
Total links	4,191	441	1,691
Accuracy	99.55%	98.87%	97.35%
Running time	Total time: 27 minutes. Including 25.9 minutes in BLAT alignment and 1 minute for PEP_scaffolder	Total time: 7,617 minutes, including 344 minutes in Augustus prediction and 7,273 in ESPRIT	Total time: 47,622 minutes, including 15,670 minutes for tblastn, 31,850 minutes for GeneWise, and 102 minutes for SWiPS.

Reference

http://genome.ucsc.edu/FAQ/FAQblat.html#blat4.

Kent, W.J. (2002) BLAT—The BLAST-Like Alignment Tool, *Genome Research*, **12**, 656-664.

Salzberg, S.L.*, et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms, *Genome Res*, **22**, 557-567.

Sims, D.*, et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses, *Nature reviews. Genetics*, **15**, 121-132.